Lecture-06: Reproducing Kernel Hilbert Space (RKHS)

1 Reproducing Kernel Hilbert Space (RKHS)

Lemma 1.1 (Cauchy-Schwarz inequality for PDS kernel). *Let K be a PDS kernel. Then*

$$K^2(x,x') \leqslant K(x,x)K(x',x') \text{ for all } x,x' \in \mathcal{X}.$$

Proof. We can write the following Gram matrix for samples x,x' and PDS kernel K as

$$\mathbf{K} = \begin{bmatrix} K(x,x) & K(x,x') \\ K(x',x) & K(x',x') \end{bmatrix}.$$

Since K is a PDS Kernel, the Gram matrix $\mathbf{K}$ is symmetric and positive semi-definite. In particular, $K(x,x') = K(x',x)$ and the $\det(\mathbf{K}) \geqslant 0$. Hence, the result follows. ☐

Definition 1.2. For any PDS kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, we can define a kernel evaluation map $\Phi_x : \mathcal{X} \to \mathbb{R}$ at a point $x \in \mathcal{X}$ by $\Phi_x(y) \triangleq K(x,y)$ for all $y \in \mathcal{X}$.

Definition 1.3. We can define a pre-Hilbert space $\mathbb{H}_0$ as the span of kernel evaluations at finitely many elements of $\mathcal{X}$. That is,

$$\mathbb{H}_0 \triangleq \left\{ \sum_{i\in I} a_i \Phi_{x_i} : I \text{ finite }, a \in \mathbb{R}^I, x \in \mathcal{X}^I \right\} \subseteq \mathbb{R}^{\mathcal{X}}.$$

The completion of $\mathbb{H}_0$ is a complete Hilbert space denoted by $\mathbb{H}$.

Theorem 1.4 (RKHS). *Let $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a PDS kernel. Then, there exists a Hilbert space $\mathbb{H}$ and a mapping $\Phi : \mathcal{X} \to \mathbb{H}$ such that for all $x,x' \in \mathcal{X}$,*

$$K(x,x') = \left\langle \Phi(x), \Phi(x') \right\rangle_{\mathbb{H}}.$$

Furthermore, $\mathbb{H}$ has the following reproducing property, for all $h \in \mathbb{H}$ and $x \in \mathcal{X}$,

$$h(x) = \langle (h(\cdot), K(x,\cdot) \rangle_{\mathbb{H}}.$$

The Hilbert space $\mathbb{H}$ is called the RKHS associated with the kernel K.

Remark 1. We make the following observations from the Theorem statement.

1. The Hilbert space $\mathbb{H} \subseteq \mathbb{R}^{\mathcal{X}}$.
2. For any $x \in \mathcal{X}$, we have $K(x,\cdot) \in \mathbb{H}$.

Proof. For any $x \in \mathcal{X}$, define $\Phi_x : \mathcal{X} \to \mathbb{R}$ such that $\Phi_x(x') = K(x,x')$. Then, we define a map $\langle \cdot,\cdot \rangle : \mathbb{H}_0 \times \mathbb{H}_0 \to \mathbb{R}$ such that fo $f = \sum_{i\in I} a_i \Phi_{x_i}$ and $g = \sum_{j\in J} b_j \Phi_{x_j}$, we have

$$\langle f,g \rangle_{\mathbb{H}_0} \triangleq \sum_{i\in I}\sum_{j\in J} a_i b_j K(x_i,x_j) = \sum_{j\in J} b_j f(x_j) = \sum_{i\in I} a_i g(x_i).$$

We can verify that the $\langle \cdot,\cdot \rangle : \mathbb{H}_0 \times \mathbb{H}_0 \to \mathbb{R}$ has the follow properties.

1. **Symmetry**: By definition, $\langle \cdot,\cdot \rangle$ is symmetric.
2. **Bilinearity**: $\langle \cdot,\cdot \rangle$ is bilinear. Can you show that $\langle \alpha f + \beta h, g \rangle = \alpha \langle f,g \rangle + \beta \langle f,g \rangle$?
3. **Positive semi-definiteness**: For any $f \in \mathbb{H}_0$, we have $f = \sum_{i\in I} a_i \Phi_{x_i}$ and since the Gram matrix $\mathbf{K}$ is symmetric and positive semidefinite for kernel K and samples $S = (x_i : i \in I)$, we have

$$\langle f,f \rangle = \sum_{i\in I}\sum_{j\in I} a_i a_j K(x_i,x_j) = a^T \mathbf{K} a \geqslant 0.$$

4. **Reproducing property:** Let $f \in \mathbb{H}_0$ and $f = \sum_{i\in I} a_i \Phi_{x_i}$. Then,
$$\langle f, \Phi_x \rangle = \sum_{i\in I} a_i K(x_i, x) = \sum_{i\in I} a_i \Phi_{x_i}(x) = f(x).$$
5. **Definiteness**: We will show that for any $f \in \mathbb{H}_0$ and $x \in \mathcal{X}$, we have bounded $f(x)$. From the reproducing property, it suffices to show that $\langle f, \Phi_x \rangle^2 \leqslant \langle f, f \rangle \langle \Phi_x, \Phi_x \rangle$ for any $x \in \mathcal{X}$. Can you show that $\langle \cdot, \cdot \rangle$ is a PDS kernel? Then the result will follow from Lemma 1.1.

From properties $1, 2, 3, 5$, it follows that $\mathbb{H}_0$ is a pre-Hilbert space which can be made complete to form the Hilbert space $\mathbb{H} = \overline{\mathbb{H}}_0$, where $\mathbb{H}_0$ is dense in $\mathbb{H}$. This Hilbert space $\mathbb{H}$ is the RKHS associated with the kernel K. □

1.1 Representer theorem

Observe that modulo the offset b, the hypothesis solution of SVMs can be written as a linear combination of the functions $K(x_i, \cdot)$, where x_i is a sample point. The following theorem known as the representer theorem shows that this is in fact a general property that holds for a broad class of optimization problems, including that of SVMs with no offset.

Theorem 1.5 (Representer theorem). *Let $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a PDS kernel and $\mathbb{H}$ its corresponding RKHS. Then for any non decreasing function $G : \mathbb{R} \to \mathbb{R}$ and any loss function $L : \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$, the optimization problem*
$$\arg\min_{h\in\mathbb{H}} F(h) = \arg\min_{h\in\mathbb{H}} G(\|h\|_{\mathbb{H}}) + L(h(x_1), \dots, h(x_m)),$$
has a solution of the form $h^* = \sum_{i=1}^m \alpha_i K(x_i, \cdot)$. If G is strictly increasing, then any solution has this form.

Proof. Let $\mathbb{H}_1 = \text{span}(K(x_i, \cdot) : i \in [m])$. We can write the RKHS $\mathbb{H}$ as the direct sum of span of $\mathbb{H}_1$ and the orthogonal space $\mathbb{H}_1^\perp$, i.e. $\mathbb{H} = \mathbb{H}_1 \oplus \mathbb{H}_1^\perp$. Hence, any hypothesis $h \in \mathbb{H}$, can be written as $h = h_1 + h_1^\perp$. Since G is non-decreasing
$$G(\|h_1\|_{\mathbb{H}}) \leqslant G(\sqrt{\|h_1\|_{\mathbb{H}}^2 + \|h_1^\perp\|_{\mathbb{H}}^2}) = G(\|h\|_{\mathbb{H}}).$$
By the reproducing property, we have for all $i \in [m]$
$$h(x_i) = \langle h, K(x_i, \cdot) \rangle = \langle h_1, K(x_i, \cdot) \rangle = h_1(x_i).$$
Therefore, $L(h(x_1), \dots, h(x_m)) = L(h_1(x_1), \dots, h_1(x_m))$, and hence $F(h_1) \leqslant F(h)$. If G is strictly increasing, then $F(h_1) < F(h)$ when $\|h_1^\perp\|_{\mathbb{H}} > 0$ and any solution of the optimization problem must be in $\mathbb{H}_1$. □

2 Empirical Kernel Map

Advantages of working with kernel is that no explicit definition of a feature map Φ is needed. Following are the advantages of working with explicit feature map Φ.
(i) For primal method in various optimization problems.
(ii) To derive an approximation based on Φ.
(iii) Theoretical analysis where Φ is more convenient.

Definition 2.1 (Empirical kernel map). Given an unlabeled training sample $x \in \mathcal{X}^m$ and a PDS kernel K, the associated **empirical kernel map** $\Phi : \mathcal{X} \to \mathbb{R}^m$ is a feature mapping defined for all $y \in \mathcal{X}$ by
$$\Phi(y) = \begin{bmatrix} K(y, x_1) \\ \vdots \\ K(y, x_m) \end{bmatrix}.$$

Remark 2. The empirical kernel map evaluated at a point $y \in \mathcal{X}$ is the vector of K-similarity measure of y with each of the m training points.

Remark 3. For any $i \in [m]$, we have $\Phi(x_i) = \mathbf{K}e_i$, where e_i is the i-th unit vector. Hence, $\langle \mathbf{K}e_i, \mathbf{K}e_j \rangle = \langle e_i, \mathbf{K}^2 e_j \rangle$. That is, the kernel matrix associated with the empirical kernel map Φ is $\mathbf{K}^2$.

Definition 2.2. Let $\mathbf{K}^\dagger$ denote the pseudo-inverse of the gram matrix $\mathbf{K}$ and let $(\mathbf{K}^\dagger)^{\frac{1}{2}}$ denote the SPSD matrix whose square is $\mathbf{K}^\dagger$. We define a feature map $\Psi : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ using the empirical kernel map Φ and the matrix $(\mathbf{K}^\dagger)^{\frac{1}{2}}$ as
$$\Psi(x) = (\mathbf{K}^\dagger)^{\frac{1}{2}}, \text{ for all } x \in \mathcal{X}.$$

Remark 4. Using the identity $\mathbf{K}\mathbf{K}^\dagger\mathbf{K} = \mathbf{K}$, we see that

$$\left\langle \Psi(x_i), \Psi(x_j) \right\rangle = \left\langle (\mathbf{K}^\dagger)^{\frac{1}{2}}\Phi(x_i), (\mathbf{K}^\dagger)^{\frac{1}{2}}\Phi(x_j) \right\rangle = \left\langle \mathbf{K}e_i, \mathbf{K}^\dagger\mathbf{K}e_j \right\rangle = \left\langle e_i, \mathbf{K}e_j \right\rangle.$$

Thus, the kernel matrix associated to map Ψ is $\mathbf{K}$.

Remark 5. For the feature mapping $\Omega : \mathcal{X} \to \mathbb{R}^m$ defined by $\Omega(x) = \mathbf{K}^\dagger\Phi(x)$ for all $x \in \mathcal{X}$, we check that the

$$\left\langle \Omega(x_i), \Omega(x_j) \right\rangle = \left\langle \mathbf{K}^\dagger\Phi(x_i), \mathbf{K}^\dagger\Phi(x_j) \right\rangle = \left\langle \mathbf{K}e_i, \mathbf{K}^\dagger e_j \right\rangle = \left\langle e_i, \mathbf{K}\mathbf{K}^\dagger e_j \right\rangle.$$

Thus, the kernel matrix associated to map Ω is $\mathbf{K}\mathbf{K}^\dagger$.

3 Kernel-based algorithms

We can generalize SVMs in the input space $\mathcal{X}$ to the SVMs in the feature space $\mathbb{H}$ mapped by the feature mapping Φ. Recall that $K(y,z) = \langle \Phi(y), \Phi(z) \rangle_{\mathbb{H}}$ for all $y, z \in \mathcal{X}$, and hence the gram matrix $\mathbf{K}$ generated by the kernel map K and the unlabeled training sample $x \in \mathcal{X}^m$ suffices to describe the SVM solution completely.

Definition 3.1 (Hadamard product). We define Hadamard product of two vectors $x, y \in \mathbb{R}^m$ as $x \circ y \in \mathbb{R}^m$ such that $(x \circ y)_i = x_i y_i$ for all $i \in [m]$.

Remark 6. We can write the dual problem for non-separable training data in this high dimensional space $\mathbb{H}$ as

$$\max_{\alpha} \mathbf{1}^T\alpha - \frac{1}{2}(\alpha \circ y)^T\mathbf{K}(\alpha \circ y)$$
$$\text{subject to: } 0 \leqslant \alpha \leqslant C \text{ and } \alpha^T y = 0.$$

The solution hypothesis h can be written as $h(x) = \text{sign}\left(\sum_{i=1}^m \alpha_i y_i K(x_i, x) + b\right)$, where $b = y_i - (\alpha \circ y)^T\mathbf{K}e_i$ for all x_i such that $0 < \alpha_i < C$.